Stochastic Processes and their Applications 31 (1989) 133-149
North-Holland

ABSTRACT STOCHASTIC APPROXIMATIONS AND APPLICATIONS

Adam SHWARTZ*
Faculty of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, 32000 Israel

Nadav BERMAN
Department of Mechanical Engineering, Ben-Gurion University of the Negev, Beer-Sheva, 84105 Israel

Received 1 July 1987
Revised 8 June 1988

Results on the convergence with probability one of stochastic approximation algorithms of the form

$$\theta_{n+1} = \theta_n - \gamma_{n+1} h(\theta_n) + u_{n+1}$$

are given, where the θ's belong to some Banach space and $\{u_n\}$ is a stochastic process. Using this extension of results of Kushner and Clark [10], conditions are given for the convergence of the linear algorithm

$$K_{n+1} = K_n - \frac{1}{n} X_n \circ [K_n X_n - Y_n].$$

Several applications of the linear algorithm to problems of identification of (possibly distributed) systems and optimization are given. The applicability of these conditions is demonstrated via an example. The systems considered here are more general than those considered by Kushner and Shwartz [12].

stochastic approximation in Banach space * strong convergence * linear algorithms

1. Introduction

Recursive algorithms are used for a large number of applications: from Newton's method for finding a zero of a function, stochastic approximations for finding such a zero in the presence of "measurement noise", to various applications in estimation, adaptive control, learning systems and other fields; see e.g. Nevelson and Hasminski [18], Ljung and Soderstrom [15], Chen [2], Kumar and Varaiya [7], Herkenrath, Kalin and Vogel [6] and references therein. Consider the following form of the algorithm ([10, 16]):

$$\theta_{n+1} = \theta_n - \gamma_{n+1} h(\theta_n) + u_{n+1}. \tag{1.1}$$

Following the introduction of the O.D.E. (Ordinary Differential Equation) method by Ljung [13], a compactness method was introduced by Kushner and Clark [10]

* Research supported in part by grant No. 85-00306 from the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel, and in part by Technion V.P.R. fund-Philipson fund for Electrical Power.

for obtaining the w.p. 1 convergence of stochastic approximation algorithms to a stable point of an associated O.D.E. Weak convergence (in the probabilistic sense) results giving more precise characterization were obtained by Kushner [8], Kushner and Shwartz [11]. The algorithm with "linear dynamics" (see below) is the well-known Widrow algorithm, and has been extensively investigated; see e.g. Eweda and Macchi [4] and references therein. A martingale approach to w.p. 1 convergence was developed recently by Métivier and Priouret [17]. Some recent applications of stochastic approximations are in the areas of adaptive control of linear systems (Becker, Kumar and Wei [1]), random access communications (Hajek [5]), estimation in linear systems (Métivier and Priouret [16]), and adaptive control of Markov chains (Shwartz and Makowski [25]).

The applicability of the stochastic approximation algorithm led to attempts to generalize the algorithm from $\mathbb{R}^K$ to more abstract spaces. Such a generalization was considered by Revez [20], with applications to learning processes. He considers operators h which are bounded from below and from above by linear operators, and obtains exponential bounds on the rate of convergence (in probability). Walk [27] proves convergence when h is a bounded linear operator whose spectrum is contained in $\{\lambda \geqslant \varepsilon > 0\}$. In [26] he also gives an invariance principle (central limit theorem) when h is linear with spectrum contained in $\{\lambda \geqslant \frac{1}{2}\}$. Salov [22] obtains a convergence result when h is "stable". In these papers (except [27]) a type of "conditional independence" of the sequence $\{u_n\}$ is assumed, and ideas from Lyapunov stability play a crucial role. In Kushner and Shwartz [12] a weak convergence method was applied to a linear compact h, relaxing the independence assumption to mixing conditions. It was shown that Hilbert-space-valued stochastic approximations arise in the identification and optimization of linear systems, and the (weak convergence) limits of these algorithms were shown to satisfy a Hilbert space valued O.D.E.

In Section 2 we provide some motivation for the study of abstract stochastic approximation algorithms, by deriving some applications of such algorithms with "linear dynamics". We then give a general formulation of such an algorithm for Hilbert-space valued processes. The resulting linear operator is usually compact, hence the point zero is in its spectrum. This is in fact the major stumbling block in extending the standard finite dimensional methods, and prevents the application of the simpler method of Walk [27]: see the last remark in Section 4. In order to prove the convergence we extend the compactness ideas of Kushner and Clark [10], to obtain conditions for convergence of the non-linear algorithm (1) in Banach spaces. Using this extension it becomes possible to relax some of the independence and spectral assumptions mentioned above, and obtain new conditions for convergence. Following a statement of the basic theorems, Section 2 closes with an example where the hypotheses of the linear case are verified.

In Section 3 the convergence theorems for the general, and then the linear case are proved. The proof for the linear case relies on the result for the general algorithm, and on several estimates which are derived in Section 4.

Notation

We deal with stochastic processes on $\{\Omega, \mathscr{F}, P\}$ (which we assume to be complete), taking their values in Hilbert spaces H_i. For random variables $X \in H_1$ and $Y \in H_2$ we denote by $X \circ Y$ the tensor product between X and Y. Using the identification of H with H^*, we shall view $X \circ Y$ as an operator from H_1 to H_2, i.e. a member of $\mathscr{L}(H_1, H_2)$, or as a bounded linear functional on $H_1 \otimes H_2$, which we denote by $X \circ Y \in \mathscr{L}(H_1 \otimes H_2)$. Denote by $\mathscr{H}_w$ the space $\mathscr{L}(H_1, H_2)$ when equipped with the weak topology of $\mathscr{L}(H_1 \otimes H_2)$. Thus, with some abuse of standard terminology, a sequence $\{A_n\}$ in $\mathscr{L}(H_1, H_2)$ is said to converge in the *weak operator topology* to an operator A if $((A_n - A)x, y)_{H_2} \to 0$ for all $x \in H_1$ and $y \in H_2$.

2. Motivation, applications and the main results

In this section we consider the applications of the following recursive algorithm:

$$K_{n+1} = K_n - \frac{1}{n} X_n \circ [K_n X_n - Y_n] \tag{2.1}$$

where X_n and Y_n take values in the Hilbert spaces H_1 and H_2 respectively, and $K_n : H_1 \to H_2$ are operators in $\mathscr{L}(H_1, H_2)$. In Section 4 we prove that under the appropriate assumptions on $\{X_n\}$ and $\{Y_n\}$, this algorithm converges, and the limit is characterized. Below we motivate the study of this algorithm by giving several applications where such abstract stochastic approximation algorithms arise. For the most part, these are identification problems where a "system" is to be identified from observations of "input" and "output", or a "best" linear approximation of some operator is sought.

Motivation and the linear algorithm

Given a record of data $\{X_n\}_n, \{Y_n\}_n$ considered to be sequences of random variables that take values in Hilbert spaces H_1 and H_2 respectively, assume X_n and Y_n are related by the model

$$Y_n = KX_n + \psi_n \tag{2.2}$$

where $\{\psi_n\}_n$ (the "observation noise") is an H_2-valued sequence of r.v.'s which are independent of $\{X_n\}$, and $K : H_1 \to H_2$ is a bounded linear operator, i.e. $K \in \mathscr{L}(H_1, H_2)$. In the identification problem K is unknown and the problem is to find an approximation of K from the available data. As in the finite dimensional case one may try to find the K satisfying the minimum variance criterion, i.e.

$$\min_{K \in \mathscr{L}(H_1, H_2)} E(Y_n - KX_n, h)^2_{H_2} \quad \forall h \in H_2.$$

Under standard stationarity and moment assumptions, the optimal solution satisfies

$$KR = f, \tag{2.3a}$$

$$R = E(X_n \circ X_n), \quad f = E(X_n \circ Y_n), \tag{2.3b}$$

i.e. R is the unique symmetric operator in $\mathscr{L}(H_1)$ and f the unique operator in $\mathscr{L}(H_1, H_2)$ defined by $E(X_n, h_1)_{H_1}(X_n, h_2)_{H_1} = (Rh_1, h_2)_{H_1}$, $(fh_1, h_2)_{H_2} = E(X_n, h_1)_{H_1} \cdot (Y_n, h_2)_{H_2}$, $h_1 \in H_1$, $h_2 \in H_2$. Thus equation (2.3) can be written as

$$E\{K(X_n \circ X_n) - X_n \circ Y_n\} = 0. \tag{2.4}$$

If H_1 and H_2 are separable so that Borel measurability is equivalent to Strong measurability and if $E\|X_n\|^2 < \infty$ and $E\|Y_n\|^2 < \infty$ then it is not difficult to show that the operators $X_n \circ X_n$ in $\mathscr{L}(H_1, H_1)$ and $X_n \circ Y_n$ in $\mathscr{L}(H_1, H_2)$ are both strongly measurable random variables and Bochner integrable i.e.

$$E(X_n \circ X_n) = \int_\Omega (X_n \circ X_n)\,\mathrm{d}P, \quad E(X_n \circ Y_n) = \int_\Omega (X_n \circ Y_n)\,\mathrm{d}P$$

are both well defined as Bochner integrals, so that (2.4) is well defined in the strong sense.

Motivated by the Robbins–Monro procedure as applied in the finite dimensional case the obvious infinite dimensional analog for the recursive estimation of K may be formulated as

$$K_{n+1} = K_n - \frac{1}{n}[K_n(X_n \circ X_n) - X_n \circ Y_n]. \tag{2.5}$$

Since by definition $K_n(X_n \circ X_n) = X_n \circ (K_n X_n)$, (2.5) may be written in the more familiar form

$$K_{n+1} = K_n - \frac{1}{n} X_n \circ [K_n X_n - Y_n]. \tag{2.6}$$

It will be convenient from the point of view of analysis to deal with the adjoint counterpart of (2.5);

$$U_{n+1} = U_n - \frac{1}{n}[(X_n \circ X_n)U_n - Y_n \circ X_n] \triangleq U_n - \frac{1}{n}[R_n U_n - f_n^*]. \tag{2.7}$$

Obviously, provided $U_0 = K_0^*$ one has $U_n = K_n^*\ \forall n$. From this it is clear that convergence properties of (2.5) and (2.7) are equivalent. In either case (2.5) or (2.7), the basic process is Hilbert valued, and the resulting stochastic approximations K_n and U_n take values in a Banach space.

Examples

A large number of Engineering problems are characterized by models that are described by (2.2).

Example 1. An image is described by the (real valued) function $f(x, y)$. The image is processed by a system (which may be an optical channel, or an image processing system) whose effect is given by

$$g(x, y) = \iint_{-\infty}^{\infty} h(x_1, y_1) f(x - x_1, y - y_1)\, \mathrm{d}x_1\, \mathrm{d}y_1$$

where $g(x, y)$ describes the available picture, and $h(x, y)$ is the impulse response of the processing system. It is assumed that $\iint_{-\infty}^{\infty} |h(x, y)|^2\, \mathrm{d}x\, \mathrm{d}y < \infty$. An algorithm to identify h, which uses samples of the picture along one axis (this may be the time axis) may be obtained as follows. Define X_n by $(X_n)(s, u) = f(s, n - u)$. Since f is a two parameter stochastic process, X_n may be considered an H_1-valued random variable where $H_1 = L_2(\mathbb{R}^2, \mathbb{R})$. If Y_n is defined by $(Y_n)(x) = g(x, n)$, then $Y_n = KX_n$ where

$$(Y_n)(x) = \iint_{-\infty}^{\infty} h(x_1, y_1) f(x - x_1; n - y_1)\, \mathrm{d}x_1\, \mathrm{d}y_1 \triangleq (KX_n)(x).$$

Thus $Y \in H_2$ where $H_2 = L_2(\mathbb{R}, \mathbb{R})$, provided appropriate conditions are satisfied.

Example 2. Stochastic approximations with linear dynamics generally arise when the underlying system is linear. Suppose, for example that $\{X(t)\}_t$ is a stochastic process that takes values in a Hilbert space H, and assume the input–output relation is given by

$$Y(t) = \int_0^1 h(\tau) X(t - \tau)\, \mathrm{d}\tau \tag{2.8}$$

where $h(\tau) \in \mathscr{L}(H, H)$ for $\tau \in [0, 1]$, $h(\tau)$ is strongly measurable (Lebesgue) and $\int_0^1 \|h(\tau)\|_{HH}^2\, \mathrm{d}\tau < \infty$. Suppose the observed output is obtained by sampling the output with additive noise ψ_n, i.e.

$$Y_n = \int_0^1 h(\tau) X(n - \tau)\, \mathrm{d}\tau + \psi_n \tag{2.9}$$

where $\{\psi_n\}_n$ are H-valued r.v.'s. Now define $(X_n)(\tau) \triangleq X(n - \tau)$. Provided some assumptions on the processes involved are enforced it is possible to view $\{Y_n\}$, $\{X_n\}$, $\{\psi_n\}$ as discrete time stochastic processes in appropriate Hilbert spaces; $X_n \in H_1 = L_2([0, 1], H)$, $Y_n \in H_2 = H$ and $\psi_n \in H_2$. So one arrives at (2.1) where for f in H_1, $Kf = \int_0^1 h(\tau) f(\tau)\, \mathrm{d}\tau$ (A Bochner integral). Thus, $K : H_1 \to H_2$ is bounded since $\int_0^1 \|h(\tau)\|_{HH}^2\, \mathrm{d}\tau < \infty$, $K^* : H \to L_2([0, 1], H)$ is given by $(K^* g)(t) = h^*(t) g$. Thus the conjugate stochastic approximation algorithm for the identification of h is given by

$$h_{n+1}^*(t) = h_n^*(t) - \frac{1}{n}\left[\int_0^1 h_n(\tau) X(n - \tau)\, \mathrm{d}\tau - Y_n\right] \circ X(n - t), \tag{2.10}$$

$$U_{n+1}(t) = U_n(t) - \frac{1}{n} X(n - t) \circ \left[\int_0^1 U_n(\tau) X(n - \tau)\, \mathrm{d}\tau - Y_n\right]. \tag{2.11}$$

It is clear that $U_0(\tau) = h_0(\tau)$ on $[0, 1]$ implies $U_n(t) = h_n(t)$ for all n.

When $H = \mathbb{R}$, (2.11) reduces to the algorithm considered by Kushner and Shwartz [12].

Example 3. Consider the following approximation problem. The output of a system is given by

$$\frac{\mathrm{d}U}{\mathrm{d}t} = AU + X(t), \quad U(0) = 0 \quad \text{or}$$

$$U(t) = \int_0^t T(t-s)X(s)\,\mathrm{d}s = \int_0^t T(s)X(t-s)\,\mathrm{d}s$$

where A is the infinitesimal generator of a strongly continuous semigroup $T(t)$ defined on a Hilbert space H and X is a stochastic process. Assuming $T(t)$ is stable, i.e. $\|T(t)x\| \to 0$ $\forall x \in H$ as $t \to \infty$, it makes sense to approximate $T(t)$ by an impulse response $h(t)$ with a compact support, so as to fit the data $\{X_n : (X_n)(t) = X(n-t)$ and $Y_n = U(n)\}$ to the model: $Y_n = \int_0^1 h(\tau)X(n-\tau)\,\mathrm{d}\tau \triangleq KX_n$ where $h(\tau)$ is an $\mathscr{L}(H)$ valued operator.

Preview of the results

Let H_1 and H_2 be separable Hilbert spaces and let $\{X_n\}$ ($\{Y_n\}$) be an H_1-valued (H_2-valued, resp.) stochastic process. For the linear algorithm, the following conditions are enforced:

L1. $(1/n)\sum_{k=1}^n X_k \circ X_k \to R$, a symmetric trace class operator a.e. in the uniform operator topology.

L2. $(1/n)\sum_{k=1}^n \|X_n\|_{H_1}^2 \to \operatorname{trace} R$ a.e. as $n \to \infty$.

L3. $(1/n)\sum_{k=1}^n X_k \circ Y_k \to E(X_1 \circ Y_1) \triangleq f$ a.e. in the uniform operator topology.

L4. $KR = f$ has a unique solution $K \in \mathscr{L}(H_1, H_2)$.

L5. For some $\alpha > 0$, $(1/n)\sum_{k=1}^n k^\alpha(K(X_k \circ X_k) - X_k \circ Y_k) \to 0$ a.e. in the uniform operator topology.

L6. The range of R is dense in H_1.

Theorem 2.1. *Assume* L1–L6. *Then $K_n \to K$ with probability one in the weak operator topology.*

This result is well known in the finite-dimensional case, under various boundedness and mixing conditions on the sequences $\{X_n\}$ and $\{Y_n\}$ [15]. The proof for the infinite dimensional case is given in Section 3. It is based on the following extension of the Kushner and Clark convergence theorem [10] from $\mathbb{R}^n$ to the abstract setting required here. Consider the algorithm (1.1), under the hypotheses (A1)–(A6) of Section 3. Let θ_* denote a stable point of the O.D.E.

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -h(\theta(t)) \tag{2.12}$$

and let C be a compact set, contained in the domain of attraction of θ_*.

Theorem 2.2. *Assume* A1–A6. *If $\theta_n \in C$ infinitely often, then the $\theta_n \to \theta_*$.*

Before embarking on the proofs, let us illustrate that the conditions of Theorem 2.1 are reasonable.

Example. Let $\{X(t)\}_t$, $t \geq 0$ be a real valued stochastic process defined on $(\Omega, \mathcal{F}, P)$ and such that

B1. $\{X(t)\}_t$ is stationary (in the strict sense) and ergodic.

B2. $(1/T)\int_0^T X(t)\,\mathrm{d}t \to EX(0) = 0$ P-a.e. as $T \to \infty$.

B3. $(1/T)\int_0^T X(t)^2\,\mathrm{d}t \to E|X_0|^2 < \infty$ P-a.e. as $T \to \infty$.

B4. $X(t, \omega)$ is measurable as a function that takes $(R \times \Omega, \mathcal{B}(R) \times \mathcal{F})$ into $(R, \mathcal{B}(R))$.

B5. $\{X(t)\}$ is Gaussian.

Under B5, H_1 can (and will) be chosen so that the null space of the covariance operator is empty.

Consider now Example 2 where it is required to identify the impulse response $h(\tau)$ of the system (2.8) and where $h \in L_2[0, 1]$. The observations are given by (2.9) where $\{\psi_n\}_n$ are i.i.d. r.v.'s with $E(\psi_n) = 0$, $E(\psi_n^2) = \sigma^2$ and $\{\psi_n\}_n$ is independent of (the sigma-algebras generated by the process) $\{X(t)\}_t$.

Let $H_1 = L_2[0, 1]$, $H_2 = \mathbb{R}$, and let $X_n \in L_2[0, 1]$ be defined by $(X_n)(t) = X(n - t)$, $t \in [0, 1]$. Then X_n is an $L_2[0, 1]$-valued random variable on $(\Omega, \mathcal{F}, P)$, and Y_n satisfies (2.2), where K is simply an element of H_1 (since H_1 is real, H_1 is identified with H_1^*). The corresponding stochastic approximation algorithm is given by (cf. Kushner and Shwartz [12])

$$h_{n+1}(\tau) = h_n(\tau) - \frac{1}{n} X(n - \tau)\left[\int_0^1 h_n(\sigma) X(n - \sigma)\,\mathrm{d}\sigma - Y_n\right], \quad \tau \in [0, 1]. \tag{2.13}$$

Applying Theorem 2.1 and provided that assumptions L1–L5 are all satisfied, one concludes that the algorithm converges in the weak topology of $L_2[0, 1]$ to the unique solution of the linear equation $KR = f$, where in this case R is given by $R(u - s) = E[X(u)X(s)]$, and $f = \int_0^1 h(\tau) R(t - \tau)\,\mathrm{d}\tau$.

The lemmas below verify that B1–B5 imply L1 to L5, so that convergence a.s. follows from Theorem 2.1.

Lemma 2.3. *Under the hypotheses* B1–B5 *on* $\{X(t)\}_t$, L1 *is satisfied.*

Proof. First recall that in this case $X_n \circ X_n$ is characterized by

$$\begin{aligned}((X_n \circ X_n) f_1, f_2)_{H_1} &= (X_n, f_1)_{H_1} (X_n, f_2)_{H_1} \\ &= \int_0^1 X(n - s) f_1(s)\,\mathrm{d}s \int_0^1 X(n - s) f_2(s)\,\mathrm{d}s, \quad f_1, f_2 \in H_1.\end{aligned}$$

Let $\mathscr{H}$ be the subspace of $\mathscr{L}(H_1)$ consisting of the compact operators in $\mathscr{L}(H_1)$ equipped with the topology inherited by the uniform operator topology on $\mathscr{L}(H_1)$. With this topology $\mathscr{H}$ is a separable Banach space. The function $g: H_1 \to \mathscr{H}$ defined by $g(h) \triangleq h \circ h$, $h \in H_1$ is continuous, and $\|g(h)\|_{\mathscr{H}} = \|h\|^2_{H_1}$. Let $\bar{X}_n(\omega) \triangleq g(X_n(\omega)) = X_n(\omega) \circ X_n(\omega)$. Then $\bar{X}_n : (\Omega, \mathscr{F}) \to (\mathscr{H}, \mathscr{B})$ is strongly measurable, where $\mathscr{B}$ is the Borel σ-algebra generated by the strong topology on $\mathscr{H}$, and $\{\bar{X}_n\}_n$ is stationary. Let μ be the measure induced by P on $(\mathscr{H}^\infty, \mathscr{B}^\infty)$. Define $\{Y_n\}_n$ to be the coordinate projections, i.e. $Y_n(h) = h_n$ where $h = (h_1, h_2, \cdots)$. Let T be the shift transformation on $\mathscr{H}^\infty$, that is $T(h_1, h_2, \cdots) = (h_2, h_3, \cdots)$. By the stationarity of $\{\bar{X}_n\}_n$, μ is invariant under T. Also

$$\int_{\mathscr{H}^\infty} \|Y_1(x)\|^2_{\mathscr{H}} \, \mathrm{d}\mu(x) = \int_\Omega \|X_1\|^2_{H_1} \, \mathrm{d}P = E \int_0^1 [X(1-s)]^2 \, \mathrm{d}s = E|X(0)|^2 < \infty. \quad (2.14)$$

Since $\mathscr{H}$ is separable, the ergodic theorem implies (Parthasarathy [19, Theorem 9.4])

$$\lim_{n\to\infty} \|S_n - E(Y_1 | \mathscr{F}_0)\|_{\mathscr{H}} = 0 \quad \mu - \text{a.e.}$$

where $\mathscr{F}_0$ denotes the T-invariant sets in $\mathscr{B}^\infty$, $S_n = (1/n)\sum_{k=1}^n Y_k$, and $E(Y_1|\mathscr{F}_0)$ is the conditional expectation in $\mathscr{H}$ (see, e.g. Scalora [23]). This means that for P almost all $\omega \in \Omega$ the sequence

$$\{\bar{S}_n(\omega)\}_n \triangleq \left\{\frac{1}{n}\sum_{k=1}^n \bar{X}_n(\omega)\right\}_n = \left\{\frac{1}{n}\sum_{k=1}^n X_n(\omega) \circ X_n(\omega)\right\}_n$$

is Cauchy, so there exists a compact symmetric $\mathscr{L}(H_1)$-valued random variable $\bar{R}$ such that $\|\bar{S}_n(\omega) - \bar{R}(\omega)\|_{\mathscr{L}(H_1)} \to 0$ as $n \to \infty$. To prove that $\bar{R}(\omega) = R$ P-a.e. it suffices to show that $\{\bar{S}_n\}_n$ converges to R a.e. in the weak operator topology a.e., that is $(\bar{S}_n g_1, g_2)_{H_1} \to (Rg_1, g_2)_{H_1}$ for all g_1 and g_2 in H_1. Now,

$$\begin{aligned}(\bar{S}_n g_1, g_2)_{H_1} &= \left(\left(\frac{1}{n}\sum_{k=1}^n X_k \circ X_k\right) g_1, g_2\right)_{H_1} = \frac{1}{n}\sum_{k=1}^n (X_k, g_1)_{H_1}(X_k, g_2)_{H_1} \\ &= \frac{1}{n}\sum_{k=1}^n \int_0^1 \int_0^1 X(k-s)X(k-\sigma)g_1(s)g_2(\sigma) \, \mathrm{d}s \, \mathrm{d}\sigma \\ &= \int_0^1 \int_0^1 \left[\frac{1}{n}\sum_{k=1}^n X(k-s)X(k-\sigma)\right] g_1(s)g_2(\sigma) \, \mathrm{d}s \, \mathrm{d}\sigma.\end{aligned}$$

To prove convergence to $\int_0^1\int_0^1 R(s-\sigma)g_1(s)g_2(\sigma)\,\mathrm{d}s\,\mathrm{d}\sigma = (Rg_1, g_2)_{H_1}$, note that by the ergodicity B1,

$$\frac{1}{n}\sum_{k=1}^n X(k-s)X(k-\sigma) \to R(s-\sigma) \quad (2.15)$$

for almost all (ω, s, σ) with respect to $[P \times (l \times l)]$. In order to prove convergence it suffices to show that for those $\omega \in \Omega$ for which (2.15) holds the sequence of random variables $\xi_n(\sigma, s) \triangleq (1/n)\sum_{k=1}^{n} X(k-s)X(k-\sigma)$ are $(l \times l)$ uniformly integrable. But

$$\begin{aligned}\iint_0^1 |\xi_n(s,\sigma)|^2 \, d\sigma \, ds &= \iint_0^1 \left[\frac{1}{n}\sum_{k=1}^{n} X(k-s)X(k-\sigma)\right]^2 d\sigma \, ds \\ &\leq \iint_0^1 \left[\frac{1}{n}\sum_{k=1}^{n} X(k-s)^2\right]\left[\frac{1}{n}\sum_{k=1}^{n} X(k-\sigma)^2\right] ds \, d\sigma \\ &= \left[\frac{1}{n}\sum_{k=1}^{n}\int_0^1 X(k-s)^2 \, ds\right]^2 \\ &= \left[\frac{1}{n}\int_0^n X(s)^2 \, ds\right]^2 \to [E|X(0)|^2]^2\end{aligned}$$

which is finite by assumption B3. This then establishes the uniform integrability, which implies L1. □

Lemma 2.4. *Under* B1–B5, L2 *holds.*

Proof

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n} \|X_k\|_{H_1}^2 &= \frac{1}{n}\sum_{k=1}^{n}\int_0^1 X(k-s)^2 \, ds \\ &= \frac{1}{n}\int_0^n X(s)^2 \, ds \to E|X(0)|^2 \quad \text{(by B3).}\end{aligned}$$

On the other hand, let $\{e_i\}_i$ be any o.n. basis for H_1. Then

$$\begin{aligned}\operatorname{trace} R &= \lim_{n\to\infty}\sum_{k=1}^{n} (Re_k, e_k)_{H_1} = \lim_{n\to\infty}\sum_{k=1}^{n} E(X_1, e_k)^2 \\ &= \lim_{n\to\infty} E\left\{\sum_{k=1}^{n} (X_1, e_k)^2\right\}.\end{aligned}$$

Since $\sum_{k=1}^{n} (X_1, e_i)_{H_1}^2 \to \|X_1\|_{H_1}^2$ as $n \to \infty$, $\sum_{k=1}^{n} (X, e_i)^2 \leq \|X_1\|_{H_1}^2$ and $E\|X_1\|_{H_1}^2 < \infty$ so that Lebesgue dominated convergence theorem applies and we conclude that trace $R = E\|X_1\|_{H_1}^2$. But $E\|X_1\|_{H_1}^2 = E\int_0^1 X(1-s)^2 \, ds$. Fubini's theorem now implies trace $R = E|X(0)|^2$ since

$$E\|X_1\|_{H_1}^2 = \int_0^1 [EX(1-s)^2] \, ds = E|X(0)|^2. \qquad \square$$

Lemma 2.5. *Under* B1–B5, L3 *holds, that is* $(1/n)\sum_{k=1}^{n} f_k \to f$ *a.e. in the strong topology of* H_1.

Proof. Note that $f_n = Z_n X_n = X_n(X_n, h)_{H_1} + \psi_n X_n$. By the first lemma $(1/n)\sum_{k=1}^{n} X_n(X_n, h)_{H_1} \to Rh$. Thus it is left to show that $(1/n)\sum_{k=1}^{n} \psi_n X_n \to 0$ a.e.

in the strong topology of H_1. Clearly $Y_n = \psi_n X_n$ is stationary. As in Lemma 2.3, $S_n \triangleq (1/n)\sum_{k=1}^{n} Y_k$ is Cauchy P-a.e. Define a sequence of real valued r.v.'s $\xi_n^y = \psi_n(X_n, y)_{H_1}$, $y \in H_1$. For each $y \in H_1$, $\{\xi_n^y\}$ is a martingale difference, and

$$\sum_{n=1}^{\infty} \frac{E|\xi_n^y|^2}{n^2} = \sigma^2 \sum_{n=1}^{\infty} \frac{E(X_n, y)_{H_1}^2}{n^2} = \sigma^2 (Ry, y)_{H_1} \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.$$

Using now Chow's result [3] it follows that $(1/n)\sum_{k=1}^{n} \xi_n^y \to 0$ $\forall y \in H_1$ a.e. Since H_1 is separable, there exists $A \subset \Omega$ with $P(A) = 1$ such that $(1/n)\sum_{k=1}^{n} \xi_n^y \to 0$ on A for all $y \in H_1$. □

Lemma 2.6. *Under* B1–B5, *$Rh = f$ has a unique solution $h \in H$, so that* L4 *holds.*

Proof. By definition (cf. (2.2)–(2.3)), under L1–L3, $f \in \text{Range}(R)$, hence there exists a solution $h \in H_1$. Since $\{X(t)\}_t$ is Gaussian the range of R is dense in H_1, and so the solution is unique. □

Remark. Assumption B5 is used only in Lemma 2.6, and it is clear from the proof that under the weaker assumption that the range of R is dense the conclusions still hold.

Lemma 2.7. *Under* B1–B5 *there exists $\alpha > 0$ such that $(1/n)\sum_{k=1}^{n} k^{\alpha}(R_k h - f_k) \to 0$ so that* L5 *holds.*

Proof. Let $\alpha < \frac{1}{2}$ and $\xi_n^y = n^{\alpha}\psi_n(X_n, y)_{H_1}$, $y \in H_1$. Then ξ_n^y is a martingale difference and

$$\sum_{n=1}^{\infty} \frac{E|\xi_n^y|^2}{n^2} \leq \sigma^2 (Ry, y)_{H_1} \sum_{n=1}^{\infty} \frac{1}{n^{2-2\alpha}} < \infty.$$

Therefore as in Lemma 2.5, $(1/n)\sum_{k=1}^{n} \xi_n^y \to 0$, P-a.e. for all $y \in H_1$. The proof concludes exactly as in Lemma 2.4. □

3. The convergence theorems

In Kushner and Shwartz [12], it became clear that the norm topology of a Hilbert space is too strong to obtain convergence. In fact, even under the weak topology they did not prove convergence of the unconstrained stochastic approximation algorithm, but only characterize the possible limits. On the other hand, an extension of the standard methods of the linear algorithms to the infinite-dimensional case is possible only if the linear operator is bounded below [27]. This is not the case in most applications, and it therefore seems necessary to combine the compactness ideas of [10] with the approach to proving boundedness. In this section we formulate the non-linear problem, covering the Banach space case and also some weaker topologies in Banach spaces. The extension of the compactness methods to the abstract setting allows treating some cases which are more general than those of [12].

Let B be a Banach space. Fix $\theta_0 \in B$, define $\{\theta_n\}$ by (1.1) where $u_i \in B$, $h: B \to B$ and assume:

A1. $\gamma_i > 0$, $\sum_{i=0}^{n} \gamma_i \to \infty$, $\gamma_i \to 0$.

Interpolating $\{u_i\}$, $\{\theta_i\}$, we obtain B-valued functions in the following way [10]; define $t_n = \sum_{i=0}^{n} \gamma_i$ and $m(n, T) = \max\{j: \sum_{i=n}^{j} \gamma_i \leq T\}$. Define the function $\theta^0(t)$ by $\theta^0(t) = \theta_n$ on $[t_{n-1}, t_n)$ and $\theta^0(t) = \theta_0$ for $t \leq 0$. The shifted functions are $\bar{\theta}^n(t) = \theta^0(t + t_n)$.

Let $\theta^n(t)$ be the linear interpolation-version of $\bar{\theta}^n(t)$, i.e. $\theta^n(t) = \bar{\theta}^n(t)$ if t is a jump point of $\bar{\theta}^n(t)$, and $\theta^n(t)$ is linear between such points. Set $\bar{U}^0(t) = \sum_{i=0}^{n} u_i$ on $[t_{n-1}, t_n)$, $\bar{U}^0(t) = 0$ for $t < 0$, and let $U^0(t)$ be a linear interpolation as above. Finally, define $U^n(t) = U^0(t + t_n) - U^0(t_n)$.

Let $F \subset B$ be a convex set containing the origin, which is a compact metric space under some invariant metric $r(x, y) \triangleq r(x - y)$ (in particular, we have in mind the norm topology and the weak topology. In either case, if F is compact, so is $\overline{\text{co}}\{F \cup \{0\}\}$, so that the assumption that F is convex and contains the origin is without loss of generality). Denote by $C_F \triangleq C_F(-\infty, \infty)$ the space of F valued continuous functions with the topology of uniform convergence on bounded intervals. Assume:

A2. $\theta^n(t) \in F$ for all n and t.

A3. $h: F \to F$ continuous and $\text{co}(h(F))$ is bounded.

A4. For each T, $\lim_{t\to\infty} \sup_{|s| \leq T} r(U^0(t+s) - U^0(t)) = 0$ and $\theta^n(t) - U^n(t) \in F$ for $|t| \leq T$ provided n is large.

A5. For all F valued, right continuous functions η with left limit, the integral below is well defined for all t such the integral remains in F, and the following holds (uniformly on bounded intervals):

$$\eta_n \to \eta \quad \text{implies} \quad \int_0^t \eta_n(s)\, \mathrm{d}s \to \int_0^t \eta(s)\, \mathrm{d}s.$$

Remarks. The hypotheses above are exactly those of the finite dimensional case, except that continuity of h has been relaxed in some cases; see, e.g. Kushner and Shwartz [11], Kushner [9] and references therein.

A2 is equivalent to the requirement that $\{\theta_n\} \in F$, since F is convex. The continuity of h and compactness of F imply that $\{h(F)\}$ is compact; under either the strong or the weak topology, this implies that $\text{co}\{h(F)\}$ is also compact, hence bounded. A4 requires, in particular, that for n large enough, $U^n(t) \in F$ for $|t| \leq T$. A5 clearly holds under the norm topology and under the weak topology (see e.g. Rudin [21]).

The compactness method of Kushner and Clark [10] is now extended as follows:

Theorem 3.1. *Under* A1–A5, *$\theta^n(\cdot)$ is sequentially compact in $C_F(-\infty, \infty)$. Any limit $\theta(t)$ of $\theta^n(\cdot)$ satisfies* (2.12) *with $\theta(t) \in F$, or more precisely*

$$\theta(t) = \theta(0) - \int_0^t h(\theta(s))\, \mathrm{d}s \tag{3.1}$$

The proof uses the Ascoli–Arzela theorem and the approximation of $\theta^n(\cdot)$ by $\bar{\theta}^n(\cdot)$; for details see [24]. Define $\theta_*^\varepsilon = \{x \in F: r(x-\theta_*) < \varepsilon\}$ and let $\theta(t, x)$ be a solution of (2.12) with $\theta(0, x) = \theta(0) = x$. Recall

Definition 3.2. θ_* is an asymptotically stable solution of $\mathrm{d}\theta/\mathrm{d}t = -h(\theta)$ in D if $\theta(t, y) \to \theta_*$ $\forall y \in D$, and for each ε there exists δ such that $x \in \theta_*^\delta$ implies $\theta(t, x) \in \theta_*^\varepsilon$ $\forall t > 0$.

Remarks. D is called the domain of attraction of θ_* and is usually assumed open. Since under our assumptions $\theta^0(t) \in F$, θ_*^ε is precompact, and we only assume the convergence to hold on $D \cap F$.

Assume

A6. θ_* is an asymptotically stable solution of (2.12) in D, an open set, and all solutions of (3.1) with $x = \theta(0, x) \in D$ are *continuous* in x (that is, $\forall t$, $x_i \to x$ *implies* $\theta(t, x_i) \to \theta(t, x)$).

Again, we need only consider $x \in F$, so that in A6 and definition (3.2) the hypotheses need only be satisfied on F and not on any open sets (which are not precompact in either the norm or the weak topology).

The general convergence result for the algorithm (1.1) is the following rephrasing of Theorem 2.2 of [10]:

Theorem 3.3. *Under* A1–A6, *if for some compact set* $C \subset D$, $\theta_n \in C$ *infinitely often, then* $\theta_n \to \theta_*$.

Remark. We do not assume *global* stability of (2.12), but rather stability in D and $\theta_n \in C$ infinitely often. The proof is obtained by defining a shifted sequence $\theta_T^n(t) = \theta^n(t-T)$ and observing that under A1–A6 the sequences (of functions) are jointly precompact. The limits satisfy the O.D.E. with $\theta_T(T) = \theta(0)$. The assumed asymptotic stability of $\theta_T(\cdot)$ implies that $r(\theta(0), \theta_*)$ can be made arbitrarily small by choosing T large. By the assumed stability applied to $\theta(\cdot)$, $r(\theta(t), \theta_*)$ is small, uniformly in $t > 0$, and the result follows. For a different proof, see [10, 24].

This deterministic convergence result has the following obvious stochastic counterpart:

Theorem 3.4 [10]. *Let* $\Omega^* \subseteq \Omega$ *be a subset so that for each* $\omega \in \Omega^*$, *assumptions* A1–A6 *hold. Let* $\omega \in \Omega^*$ *be such that* $\theta_n(\omega) \in C(\omega)$ *infinitely often. Then* $\theta_n(\omega) \to \theta_*(\omega)$.

Lemma 4.2 establishes the (norm) boundedness of θ_n, so that they are contained in a (weakly) compact set $F = F(\omega)$. Using this and the above result, the proof of Theorem 2.1 proceeds as follows:

Proof of Theorem 2.1. In view of Theorem 3.4, it is sufficient to show that under L1–L5, A1–A6 holds with $F \subset D$, since F is compact. A1 is obviously satisfied by the sequence $\{1/n\}$. To prove A2 we need to show that θ_n lies in some (weakly) compact set (which may depend on ω). This is proved in Lemma 4.2 below. A3 follows easily since h is linear and from the proof of A2 and the properties of $\mathscr{H}_w$. A4 follows from a stronger result which is established in Lemma 4.3 below, and A5 follows immediately from the fact that for any $\lambda \in H_1 \otimes H_2$, $\lambda \int_0^t \eta_s \, ds = \int_0^t \lambda \eta_s \, ds$. Finally, A6 follows from the properties of R (see L1, L3 and L6). □

4. Estimates and proof of boundedness

Consider now the algorithm $U_{n+1} = U_n - (1/n)[R_n U_n - f_n^*]$ where $U_n \in \mathscr{L}(H_2, H_1)$, $R_n = X_n \circ X_n$ and $f_n^* = Y_n \circ X_n$. Let $\tilde{U}_n = U_n - U$ where U is the unique solution of $RU = f^*$, and R and f are as before. Rewrite the stochastic approximation algorithm as

$$\tilde{U}_{n+1} = \tilde{U}_n - \frac{1}{n} R\tilde{U}_n + \frac{1}{n}(R - R_n)\tilde{U}_n - \frac{1}{n}(R_n U - f_n^*) = \tilde{U}_n - \frac{1}{n} R\tilde{U}_n + \frac{1}{n}\varepsilon_n. \tag{4.1}$$

To satisfy the conditions for convergence as stated above one has to have

$$\lim_{n\to\infty} \left\{ \sup_{n<k\leq m(n,\tau)} r\left(\sum_{i=n+1}^{k} \frac{1}{i}\varepsilon_i \right) \right\} = 0 \quad \text{a.e.} \tag{4.2}$$

where $r(x, y) = r(x - y)$ is the invariant metric defined on the strong ball of $\mathscr{L}(H_2, H_1)$ containing F, and which is compatible with the weak operator topology. Recall (Section 2) that this topology is compatible with the following: $\{A_n\} \subset \mathscr{L}(H_2, H_1)$ converges to A iff $((A_n - A)x, y)_{H_1} \to 0$ $\forall x \in H_2, y \in H_1$. This topology is weaker than the weak topology (in the usual sense) of $\mathscr{L}(H_2, H_1)$. However for separable Hilbert spaces H_1, H_2 the strong operator ball in $\mathscr{L}(H_2, H_1)$ is compact in this weak topology and metrizable.

Thus (4.2) is equivalent to

$$\lim_{n\to\infty} \left\{ \sup_{n<k\leq m(n,\tau)} \left| \sum_{i=n+1}^{k} \frac{1}{i} (\varepsilon_i h_2, h_1)_{H_1} \right| \right\} = 0 \quad \text{for all } h_1 \in H_1 \text{ and } h_2 \in H_2.$$

Lemma 4.1. *Consider* (4.1). *If* $\tilde{U}_n$ *is bounded a.e., i.e.* $\{\|\tilde{U}_n\|_{H_2 H_1}\}_n$ *is bounded a.e., then*

$$\lim_{n\to\infty} \left\{ \sup_{n<k\leq m(n,\tau)} \left\| \sum_{i=n+1}^{k} \frac{1}{j}\varepsilon_i \right\|_{H_2 H_1} \right\} = 0 \quad a.e. \tag{4.3}$$

This is clearly stronger than (4.2). The proof is similar to the finite dimensional analogue; see [24].

Lemma 4.2. *Under* L1–L5 $\{U_n\}_n$ *is bounded P-a.e. in the uniform operator topology on* $\mathscr{L}(H_2, H_1)$.

Remark. This implies that the sequence $\{U_n\}_n$ lies in a compact set relative to the weak operator topology, which proves A2.

Proof. Here we deal with adjoint algorithm as described by (2.7). Consider the perturbed algorithm

$$U^\varepsilon_{n+1} = U^\varepsilon_n - \frac{1}{n}[R_n U^\varepsilon_n - Y_n \circ X_n - \varepsilon(U - U^\varepsilon_n)], \quad \varepsilon > 0. \tag{4.4}$$

Let $\tilde{U}^\varepsilon_n = U^\varepsilon_n - U$; then (4.4) yields

$$\tilde{U}^\varepsilon_{n+1} = U^\varepsilon_n - U - \frac{1}{n}[R_n(U^\varepsilon_n - U) + R_n U + \varepsilon(U^\varepsilon_n - U) - f^*_n]$$

$$= \tilde{U}^\varepsilon_n - \frac{1}{n} R_n \tilde{U}^\varepsilon_n - \frac{1}{n}(R_n U - f^*_n) - \frac{1}{n}\varepsilon \tilde{U}^\varepsilon_n. \tag{4.5}$$

In the same way set $\tilde{U}_n \triangleq U_n - U$ to obtain from (2.7)

$$\tilde{U}_{n+1} = \tilde{U}_n - \frac{1}{n} R_n \tilde{U}_n - \frac{1}{n}(R_n U - f^*_n). \tag{4.6}$$

Now subtracting (4.5) from (4.6) yields the equation for $Z^\varepsilon_n \triangleq \tilde{U}_n - \tilde{U}^\varepsilon_n$,

$$Z^\varepsilon_{n+1} = \tilde{U}_n - \tilde{U}^\varepsilon_n - \frac{1}{n} R_n(\tilde{U}_n - \tilde{U}^\varepsilon_n) + \frac{1}{n}\varepsilon \tilde{U}^\varepsilon_n = Z^\varepsilon_n - \frac{1}{n} R_n Z^\varepsilon_n + \varepsilon \frac{1}{n} \tilde{U}^\varepsilon_n. \tag{4.7}$$

Boundedness of the sequence $\{\tilde{U}_n\}_n$ will follow from the inequality

$$\|\tilde{U}_n\|_{H_2H_1} \leq \|\tilde{U}_n - \tilde{U}^\varepsilon_n\|_{H_2H_1} + \|\tilde{U}^\varepsilon_n\|_{H_2H_1} = \|Z^\varepsilon_n\|_{H_2H_1} + \|\tilde{U}^\varepsilon_n\|_{H_2H_1}. \tag{4.8}$$

In Lemma 3 and Theorem 4 below we show that $\{Z^\varepsilon_n\}_n$ and $\{\tilde{U}^\varepsilon_n\}_n$ are bounded a.e. in the uniform operator topology, and the lemma follows. □

Lemma 4.3. *If* $(1/n)R_n \to 0$ *a.e. and* $\sum_{n=1}^\infty (1/n)\|\tilde{U}^\varepsilon_n\|_{H_2H_1} < \infty$ *a.e., then* $\{Z^\varepsilon_n\}_n$ *is bounded a.e.*

Proof. Since R_n is positive definite, i.e. $(R_n x, x)_{H_1} \geq 0$ for all x in H_1, it follows that

$$\left\| I - \frac{1}{n} R_n \right\|_{H_1H_1} \leq 1 \quad \forall n \geq N \quad \text{for some } N = N(\omega) > 0.$$

Hence

$$\prod_{n=1}^{\infty} \left\| I - \frac{1}{n} R_n \right\|_{H_1H_1} \leq M \quad \text{for some } M(\omega) > 0 \text{ and } n \geq N.$$

The rest of the proof is obvious. □

In order to establish the boundedness of $\{\tilde{U}_n\}_n$ it is left to verify that the hypotheses of Lemma 3 hold. The first hypothesis is established as in the finite dimensional case, and the proof is omitted. The proof of the second hypothesis proceeds by showing that the sequence $\{n^\alpha \|\tilde{U}_n^\varepsilon\|\}_n$ is bounded P-a.e. for some $\alpha > 0$. This implies that $\{\|\tilde{U}_n^\varepsilon\|_{H_2H_1}\}_n$ is bounded and thus by (4.8), $\{\tilde{U}_n\}_n$ is bounded P-a.e. with respect to the required topology. So, fix ε and define $R_n^\varepsilon \triangleq R_n + \varepsilon I = X_n \circ X_n + \varepsilon I$.

Theorem 4.4. *Let assumptions* L1-L5 *hold and write* (4.5), *in the form*

$$\tilde{U}_{n+1}^\varepsilon = \tilde{U}_n^\varepsilon - \frac{1}{n} R_n^\varepsilon \tilde{U}_n^\varepsilon + \frac{1}{n} B_n, \quad \varepsilon > 0, \tag{4.9}$$

where $B_n = -R_n U + f_n^*$. *Let* $\alpha > 0$ *be as in* L5. *Then* $n^\alpha \tilde{U}_n^\varepsilon \to 0$ *(a.e.) in the uniform operator topology in* $\mathscr{L}(H_2, H_1)$.

Proof. Multiplying both sides of (4.9) by $(n+1)^\alpha$ yields

$$(n+1)^\alpha \tilde{U}_{n+1}^\varepsilon = \frac{(n+1)^\alpha}{n^\alpha} n^\alpha \tilde{U}_n^\varepsilon - \frac{(n+1)^\alpha}{n^\alpha} \cdot \frac{1}{n} R_n^\varepsilon n^\alpha \tilde{U}_n^\varepsilon + \frac{(n+1)^\alpha}{n} B_n,$$

$$Y_{n+1} = \left(\frac{n+1}{n}\right)^\alpha \left[Y_n - \frac{1}{n} R_n^\varepsilon Y_n \right] + \frac{(n+1)^\alpha}{n} B_n, \tag{4.10}$$

where $Y_n = n^\alpha \tilde{U}_n^\varepsilon$. Since $(n+1/n)^\alpha = 1 + (\alpha/n) + \mathrm{O}(n^{-2})$,

$$Y_{n+1} = Y_n - \frac{1}{n} R_n^\varepsilon Y_n + \frac{\alpha}{n} Y_n - \frac{\alpha}{n^2} R_n^\varepsilon Y_n + \mathrm{O}(n^{-2}) Y_n$$

$$- \frac{1}{n} \mathrm{O}(n^{-2}) R_n^\varepsilon Y_n + \frac{(n+1)^\alpha}{n} B_n.$$

Since ε is arbitrary there is no loss in generality assuming $1 + \alpha > \varepsilon > \alpha$. Let $\delta = \varepsilon - \alpha$, so one can write

$$Y_{n+1} = Y_n - \frac{1}{n} R_n^\delta Y_n + \frac{1}{n^2} [-\alpha R_n^\varepsilon + n^2 \mathrm{O}(n^{-2}) I - n \mathrm{O}(n^{-2}) R_n^\varepsilon] Y_n$$

$$+ \frac{(n+1)^\alpha}{n} B_n$$

$$= Y_n - \frac{1}{n} R_n^\delta Y_n + \frac{1}{n} C_n Y_n + \frac{(n+1)^\alpha}{n} B_n, \tag{4.11}$$

where $C_n = 1/n[-\alpha R_n^\varepsilon + n^2 \mathrm{O}(n^{-2}) I - n \mathrm{O}(n^{-2}) R_n^\varepsilon]$. Summing from n to $m+1$,

$$Y_{m+1} = Y_n - \sum_{k=n}^{m} \frac{1}{k} R_k^\delta Y_k + \sum_{k=n}^{m} \frac{1}{k} C_k Y_k + \sum_{k=n}^{m} \frac{(k+1)^\alpha}{k} B_k$$

$$= (I - \tau R^\delta) Y_n + \left(\tau R^\delta - \sum_{k=n}^{m} \frac{1}{k} R_k^\delta \right) Y_n + \sum_{k=n}^{m} \frac{1}{k} R_k^\delta (Y_n - Y_k)$$

$$- \sum_{k=n}^{m} \frac{1}{k} C_k (Y_n - Y_k) + \left(\sum_{k=n}^{m} \frac{1}{k} C_k \right) Y_n + \sum_{k=n}^{m} \frac{1}{k} \bar{B}_k \tag{4.12}$$

where $\bar{B}_k = (k+1)^\alpha B_k$. Let

$$S_n^m(\tau) = h_n^m(\tau) Y_n + r_n^m(\tau) + g_n^m(\tau)$$

where

$$h_n^m(\tau) = \tau R^\delta - \sum_{k=n}^m \frac{1}{k} R_k^\delta + \sum_{k=n}^m \frac{1}{k} C_k,$$

$$r_n^m(\tau) = \sum_{k=n}^m \frac{1}{k} R_k^\delta (Y_n - Y_k) - \sum_{k=n}^m \frac{1}{k} C_k (Y_n - Y_k)$$

and

$$g_n^m(\tau) = \sum_{k=n}^m \frac{1}{k} \bar{B}_k.$$

Now write (4.12) in the form

$$Y_{m+1} = (I - \tau R^\delta) Y_n + S_n^m(\tau). \tag{4.13}$$

Since R^δ is bounded from below, the arguments of the finite dimensional case (see e.g. Métivier and Priouret [16], Ljung [14]) can be extended in a straightforward manner to yield the boundedness of $\{Y_n\}$. For details, see [24]. □

Remark. Since in L5 α can be chosen arbitrarily small, the Lemma establishes that $\{n^\alpha \tilde{U}_n^\varepsilon\}_n$ converges to zero for any $\varepsilon > \alpha > 0$ but not for $\varepsilon = 0$. In the finite dimensional case, these methods can be applied to obtain convergence for $\varepsilon = 0$ also, and so the use of Theorem 2.2 can be avoided altogether. This cannot be extended to the infinite dimensional case, where Theorem 2.2 plays a crucial role.

References

[1] A.H. Becker, Jr., P.R. Kumar and C.-Z. Wei, Adaptive control with the stochastic approximation algorithm: geometry and convergence, IEEE Trans. Auto. Control 30 (1985) 330–338.

[2] H.-F. Chen, Recursive Estimation and Control for Stochastic Systems (Wiley, New York, 1985).

[3] Y.S. Chow, Local convergence of martingales and the law of large numbers, Ann. Math. Statist. 36 (1965) 552–558.

[4] E. Eweda and O. Macchi, Convergence analysis of an adaptive linear estimation algorithm, IEEE Trans. Auto. Control 29 (1984) 119–123.

[5] B. Hajek, Stochastic Approximation Methods for Decentralized Control of Multiaccess Communications, IEEE Trans. Inf. Theory 31 (1985) 176–184.

[6] U. Herkenrath, D. Kalin and W. Vogel, Mathematical Learning Models—Theory and Applications (Springer-Verlag, New York, 1983).

[7] P.R. Kumar and P. Varaiya, Stochastic Systems; Estimation, Identification and Adaptive Control (Prentice-Hall, New Jersey, 1986).

[8] H.J. Kushner, Convergence of recursive, adaptive and identification procedures via weak convergence theory, IEEE Trans. Auto. Control 22 (1977) 921–930.

[9] H.J. Kushner, Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory (MIT Press, Cambridge, Massachusetts, 1984).

[10] H.J. Kushner and D.S. Clark, Stochastic Approximation Methods for Constrained and Unconstrained Systems (Springer Verlag, New York, 1978).
[11] H.J. Kushner and A. Shwartz, An invariant measure approach to the convergence of Stochastic Approximations with state-dependent noise, SIAM J. Control Opt. 22 (1984) 13-27.
[12] H.J. Kushner and A. Shwartz, Stochastic Approximations in Hilbert space; identification and optimization of linear continuous-parameter systems, SIAM J. Control Opt. 23 (1985) 774-793.
[13] L. Ljung, Analysis of recursive stochastic algorithms, IEEE Trans. Auto. Control 22 (1977) 551-575.
[14] L. Ljung, Analysis of stochastic gradient algorithms for linear regression problems, IEEE Trans. Info. Theory, IT-30 (1984) 151-160.
[15] L. Ljung and T. Soderstrom, Theory and Practice of Recursive Identification (MIT Press, Mass., 1983).
[16] M. Métivier and P. Priouret, Applications of a Kushner and Clark Lemma to general classes of stochastic algorithms, IEEE Trans. Info. Theory 30 (1984) 140-150.
[17] M. Métivier and P. Priouret, Théoremes de convergence presque sure pour une classe d'algorithmes stochastiques a pas decroissants, Prob. Theory Rel. Fields 74 (1987) 403-428.
[18] M.B. Nevel'son and R.Z. Has'minskii, Stochastic Approximations and Recursive Estimation, Translation of Mathematical Monographs Vol. 47 (AMS, Providence, Rhode Island, 1973).
[19] K.R. Parthasarathy, Probability Measures on Metric Spaces (Academic Press, New York, 1967).
[20] P. Revez, Robbins-Monro procedure in a Hilbert space and its application in the theory of learning systems, Studia Scientiarum Math. Hungarica 8 (1973) 391-398.
[21] W. Rudin, Functional Analysis (McGraw-Hill, New York, 1973).
[22] G.I. Salov, Stochastic approximation theorem in a Hilbert space and its application, Theor. Prob. Appl. 24 (1979) 413-419.
[23] F.S. Scalora, Abstract Martingale convergence theorems, Pacific J. Math. 11 (1961) 347-374.
[24] A. Shwartz and N. Berman, Abstract stochastic approximations and applications, EE Pub. 630, Technion, June 1987.
[25] A. Shwartz and A.M. Makowski, An optimal adaptive scheme for two competing queues with constraints, in: Analysis and Optimization of Systems, eds. A. Bensoussan and J.L. Lions (Springer-Verlag, 1986) 515-532.
[26] H. Walk, An invariance principle for the Robbins Monro process in a Hilbert space, Z. Wahr. verw. Gebiete 39 (1977) 135-150.
[27] H. Walk, Almost sure convergence of stochastic approximation processes, 16th European Meeting of Statisticians, in: Statistics and Decisions Supplement Issue No. 2 (R. Oldenbourg, Verlag, Munchen, 1985) 137-141.